

SinDDM: A Single Image Denoising Diffusion Model

Vladimir Kulikov Shahar Yadin* Matan Kleiner* Tomer Michaeli
Technion – Israel Institute of Technology, Haifa, Israel

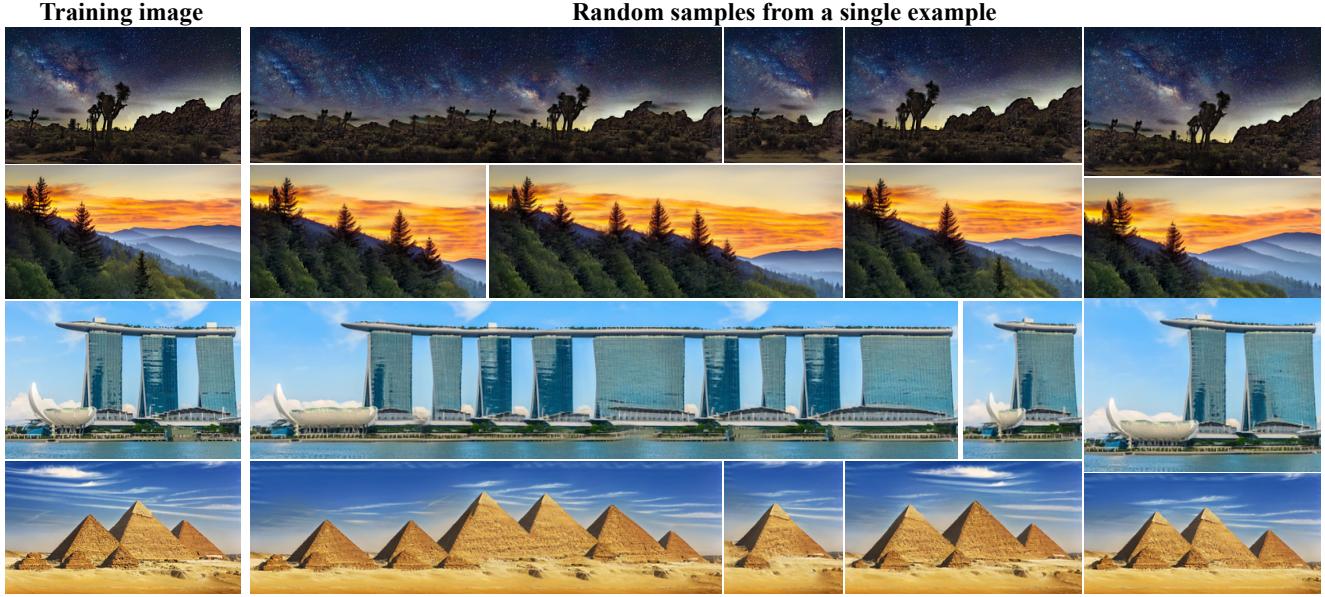


Figure 1. **Single image diffusion model.** We introduce a framework for training an unconditional denoising diffusion model (DDM) on a single image. Our single-image DDM (SinDDM) can generate novel high-quality variants of the training image at arbitrary dimensions by creating new configurations of both large objects and small-scale structures (*e.g.* the shape of the skyline in row 1 and the angles formed by the distant mountains in row 2). SinDDM can be used for many tasks, including text-guided generation from a single image (Fig. 2).

Abstract

Denoising diffusion models (DDMs) have led to staggering performance leaps in image generation, editing and restoration. However, existing DDMs use very large datasets for training. Here, we introduce a framework for training a DDM on a single image. Our method, which we coin SinDDM, learns the internal statistics of the training image by using a multi-scale diffusion process. To drive the reverse diffusion process, we use a fully-convolutional light-weight denoiser, which is conditioned on both the noise level and the scale. This architecture allows generating samples of arbitrary dimensions, in a coarse-to-fine manner. As we illustrate, SinDDM generates diverse high-quality samples, and is applicable in a wide array of tasks, including style transfer and harmonization. Furthermore, it can be easily guided by external supervision. Particularly, we demonstrate text-guided generation from a single

image using a pre-trained CLIP model. Results and code are available on the project’s [webpage](#).

1. Introduction

Image synthesis and manipulation has attracted a surge of research in recent years, leading to impressive progress in *e.g.* generative adversarial network (GAN) based methods [6] and denoising diffusion models (DDMs) [26]. State-of-the-art generative models now reach high levels of photorealism [3, 12, 23], can treat arbitrary image dimensions [2], can be used to solve a variety of image restoration and manipulation tasks [14, 20, 22], and can even be conditioned on complex text prompts [16, 18, 19, 21]. However, this impressive progress has often gone hand-in-hand with the reliance on increased amounts of training data. Unfortunately, in many cases relevant training examples are scarce.

*These authors contributed equally to this work.

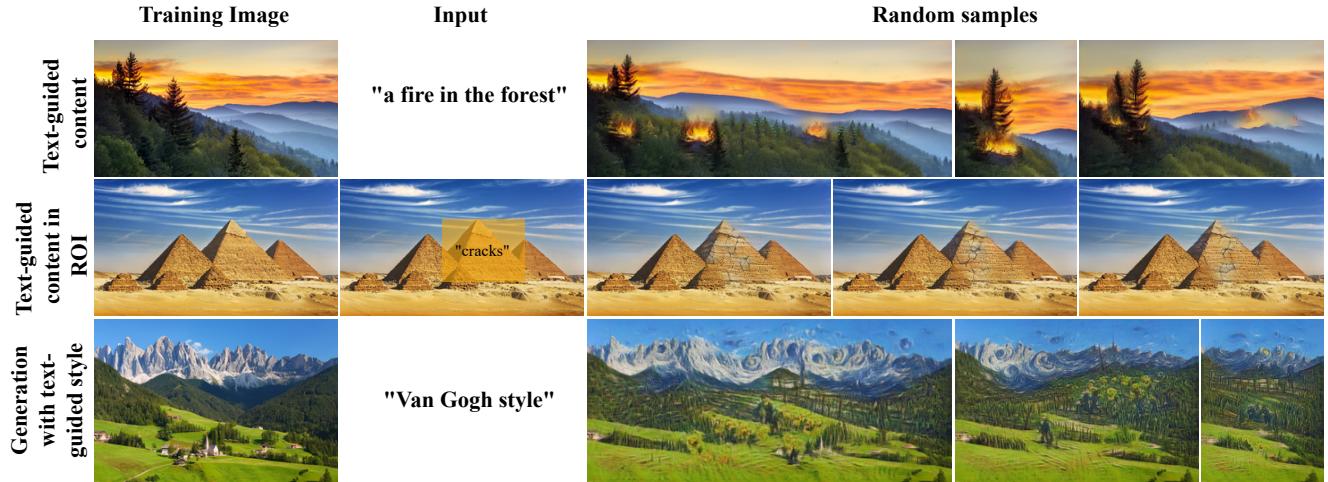


Figure 2. **Text guided generation.** SinDDM can generate images conditioned on text prompts in several different manners. We can control the contents of the generated samples across the entire image (top) or within a user-prescribed region of interest (middle). We can also control the style of the generated samples (bottom). All effects are achieved by modifying only the sampling process, without the need for any architectural changes or for training or tuning the model (see Sec. 4).

A recent line of work developed ways to learn a generative model from a single natural image. The first *unconditional* model proposed for this task was SinGAN [24]. This model uses a pyramid of patch-GANs to learn the distribution of small patches in several image scales. Once trained on a single image, SinGAN can randomly generate similar images, as well as solve a variety of tasks, including editing, style transfer and super-resolution. Follow up works improved SinGAN’s training process [10], extended it to other domains (*e.g.* audio [8], video [9], 3D shapes [29]), and used alternative learning frameworks (energy-based models [31], nearest-neighbor patch search [7], enforcement of deep feature statistics via test-time optimization [4]).

In this paper, we propose a different approach for learning a generative model from a single image. Specifically, we combine the multi-scale approach of SinGAN with the power of DDMs to devise SinDDM, a hierarchical DDM that can be trained on a single image. At the core of our method is a fully-convolutional denoiser, which we train on various scales of the image, each corrupted by various levels of noise. We take the denoiser’s receptive field to be relatively small so that it only captures the statistics of the fine details within each scale. At test time, we use this denoiser in a coarse-to-fine manner, which allows generating diverse random samples of arbitrary dimensions. As illustrated in Fig. 1, SinDDM synthesizes high quality images while exhibiting good generalization capabilities. For example, certain small structures in the skylines in row 1 and the angles of some of the mountains in row 2 do not exist in the corresponding training images, yet they look realistic.

Similarly to existing single-image generative models,

SinDDM can be used for image-manipulation tasks (see Sec. 4). However, perhaps its most appealing property is that it can be easily guided by external supervision, thus allowing to generate images in a controlled manner. For example, in Fig. 2 we demonstrate text guidance for controlling the content and style of the generated images. These effects are achieved by employing a pretrained CLIP model [17]. Other guidance options are illustrated in Sec. 4.

2. Related work

Single image generative models Single-image generative models perform image synthesis and manipulation by capturing the internal distribution of patches or deep features within a single image. In [25], the authors presented a single-image conditional GAN model for the task of image retargeting. In the context of unconditional models, SinGAN [24] is a hierarchical GAN model that can generate high quality, diverse samples based on a single training image. SinGAN’s training process was improved in [10]. Several works replaced SinGAN’s GAN framework by other techniques for learning distributions. These include energy-based models [31], nearest-neighbor patch search [4,7], and enforcement of deep-feature distributions via test-time optimization of a sliced-Wasserstein loss [4]. Here, we follow the hierarchical approach of SinGAN, but using denoising diffusion probabilistic models [12]. This enables us to generate high quality images, while supporting guided image generation as in [3].

Diffusion models First presented by Soh *et al.* [26], diffusion models sample from a distribution by reversing a gradual noising (diffusion) process. This method recently

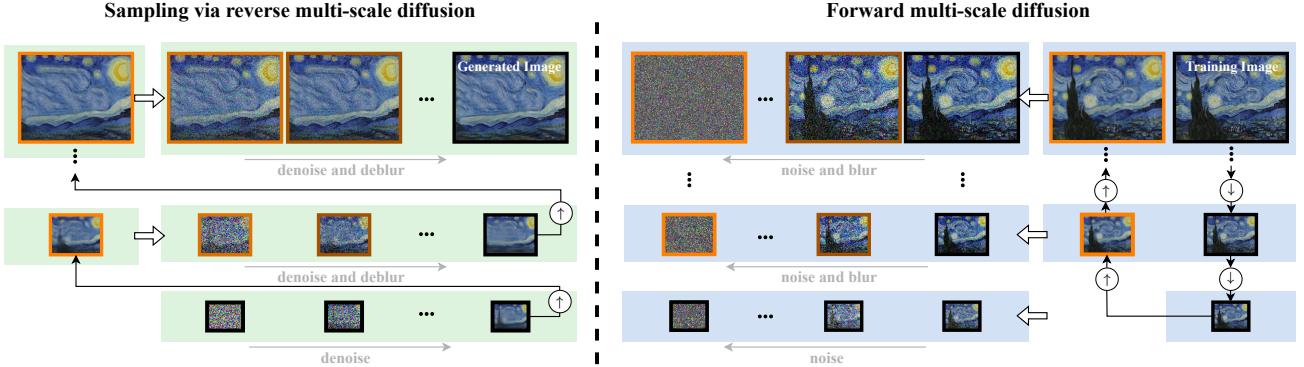


Figure 3. Multi-scale diffusion. Our forward multi-scale diffusion process (right) is constructed from down-sampled versions of the training image (black frames), as well as their blurry versions (orange frames). In each scale, we construct a sequence of images that are linear combinations of the original image in that scale, its blurry version, and noise. Sampling via the reverse multi-scale diffusion (left), starts from pure noise at the coarsest scale. In each scale, our model gradually removes the noise until reaching a clean image, which is then upsampled and combined with noise to start the process again in the next scale.

achieved impressive results in image generation [3, 12] as well as in various image restoration and manipulation tasks, including super-resolution [22], image-to-image translation [20] and image editing [14]. These works established diffusion models as the current state-of-the-art in image generation and manipulation.

Text-guided image manipulation and generation Text-guided image manipulation and generation has recently attracted considerable interest with the emergence of models like DALL-E 2 [18], stable diffusion [19] and Imagen [21]. Those models generate high quality realistic images from a text prompt, and were even extended to video generation [11]. Those techniques have also been found useful for image editing tasks, such as manipulating a set of user provided images by using text [5]. One popular way to guide image generation models by text is by using a pre-trained CLIP model [17]. In SinDDM we adopt this approach and combine CLIP’s external knowledge with our internal model to guide the image generation process by text prompts. Recently, Text2Live [1] described an approach for text-guided image editing by training on a single image. This method uses a pre-trained CLIP model to guide the generation of an edit layer that is later combined with the original image. Thus, as opposed to our goal here, Text2Live can only add details on top of the original image; it cannot change the entire scene (*e.g.* changing object configurations) or generate images whose dimensions differ from the original image.

3. Method

Our goal is to train an unconditional generative model to capture the internal statistics of structures at multiple scales within a single training image. Similarly to existing DDM

frameworks, we employ a diffusion process, which gradually turns the image into white Gaussian noise. However, here we do it in a hierarchical manner that combines both blur and noise.

3.1. Forward multi-scale diffusion

As illustrated in the right pane of Fig. 3, we start by constructing a pyramid $\{x^{N-1}, \dots, x^0\}$ with a scale factor of $r > 0$ (black frames). Each x^s is obtained by down-sampling x by r^{N-1-s} (so that x^{N-1} is the training image x itself). We also construct a blurry version of the pyramid (orange frames), $\{\tilde{x}^{N-1}, \dots, \tilde{x}^0\}$, where $\tilde{x}^0 = x^0$ and $\tilde{x}^s = (x^{s-1})^{\uparrow r}$ for every $s \geq 1$. We use those two pyramids to define a multi-scale diffusion process over $(s, t) \in \{0, \dots, N-1\} \times \{0, \dots, T\}$ as

$$x_t^s = \sqrt{\bar{\alpha}_t} (\gamma_t^s \tilde{x}^s + (1 - \gamma_t^s)x^s) + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$. As t grows from 0 to T , γ_t^s increases monotonically from 0 to 1 while $\bar{\alpha}_t$ decreases monotonically from 1 to 0 (see Appendix D for details). Therefore, as t increases, x_t^s becomes both noisier and blurrier.

3.2. Reverse multi-scale diffusion

To sample an image, we attempt to reverse the diffusion process, as shown in the left pane of Fig. 3. Specifically, we start at scale $s = 0$, where we follow the standard DDM approach (starting with random noise at $t = T$ and gradually removing noise until a clean sample is obtained at $t = 0$). We then upsample the generated image to scale $s = 1$, combine it with noise again, and run a reverse diffusion process to form a sample at this scale. The process is repeated until reaching the finest scale, $s = N - 1$.

The reverse diffusion process is driven by a single fully convolutional model, which is trained to predict x_0^s based on

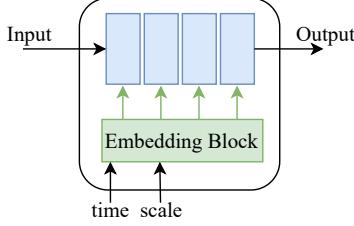


Figure 4. **SinDDM architecture.** We use a fully-convolutional model with four blocks, having a total receptive field of 35×35 . The model is conditioned on both the timestep t and the scale s .

x_t^s (in practice it predicts the noise ϵ from which we extract a prediction of x_0^s). The training procedure is summarized in Alg. 1. For sampling, we adopt the general DDIM formulation [27], as detailed in Alg. 2, where for scale $s = 0$ we use the noise variance of DDPM [12] and for $s > 0$ we use $\sigma_t^s = 0$ except when applying text-guidance, in which case we also use the DDPM scheduler. Note that for $s = 0$, $\gamma_t^s = 0$ since there is no blur to remove. More details about the γ_t^s schedule in regular sampling and in text-guided sampling are provided in appendices D and E, respectively.

As shown in Fig. 4, our model is conditioned on both the timestep t and the scale s . We found this to improve generation quality and training time compared to a separate diffusion model for each scale. Our model comprises 4 convolutional blocks, with a total receptive field of 35×35 . The number of scales is chosen such that the area covered by the receptive field is as close as possible to 40% of the area of the entire image at scale 0. In most of our experiments, this rule led to 4 or 5 scales. The small receptive field forces the model to learn the statistics of small structures and prevents memorization of the entire image. For every scale $s > 0$, we start the reverse diffusion at timestep $T[s] \leq T$, which we set such that $(1 - \bar{\alpha}_{T[s]})/\bar{\alpha}_{T[s]}$ is proportional to the MSE between x^s and \tilde{x}^s . This ensures that the amount of noise added to the upsampled image from the previous scale is proportional to the amount of missing details at that scale (see derivation in App. D). For $s = 0$, we start at $T[0] = T$.

As opposed to external DDMs, our model uses only convolutions and GeLU nonlinearities, without any self-attention or downsampling/upsampling operations. The timestep t and scale s are injected to the model using a joint embedding, similarly to the one used to inject only t in [12] (see App. B). The model has a total of 1.1×10^6 parameters and its training on a 250×200 image takes around 7 hours on an A6000 GPU. Sampling of a single image takes a few seconds. In each training iteration we sample a batch of noisy images from the same randomly chosen scale s but from several randomly chosen timesteps t . We train the model for 120,000 steps using the Adam optimizer with its default parameters (see App. C for further details).

Algorithm 1 SinDDM Training

```

1: repeat
2:    $s \sim \text{Uniform}(\{0, \dots, N - 1\})$ 
3:    $t \sim \text{Uniform}(\{0, \dots, T\})$ 
4:    $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
5:   if  $s = 0$  then
6:      $x_t^{s,\text{mix}} = x^s$ 
7:   else
8:      $x_t^{s,\text{mix}} = \gamma_t^s x^{s-1} \uparrow^r + (1 - \gamma_t^s)x^s$ 
9:   end if
10:  Update model  $\epsilon_\theta$  by taking gradient descent step on
11:     $\nabla_\theta \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_t^{s,\text{mix}} + \sqrt{1 - \bar{\alpha}_t}\epsilon, t, s) \right\|_1$ 
12: until converged

```

Algorithm 2 SinDDM Sampling

```

1: for  $s = 0, \dots, N - 1$  do
2:   if  $s = 0$  then
3:      $x_{T[0]}^0 \sim \mathcal{N}(0, \mathbf{I})$ 
4:   end if
5:   for  $t = T[s], \dots, 1$  do
6:      $x_t^{s,\text{mix}} = \frac{x_t^s - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t^s, t, s)}{\sqrt{\bar{\alpha}_t}}$ 
7:      $\hat{x}_0^s = \frac{x_t^{s,\text{mix}} - \gamma_t^s \tilde{x}^s}{1 - \gamma_t^s}$ 
8:      $x_{t-1}^{s,\text{mix}} = \gamma_{t-1}^s \tilde{x}^s + (1 - \gamma_{t-1}^s)\hat{x}_0^s$ 
9:      $z \sim \mathcal{N}(0, \mathbf{I})$ 
10:     $x_{t-1}^s = \sqrt{\bar{\alpha}_{t-1}}x_{t-1}^{s,\text{mix}}$ 
11:     $+ \sqrt{1 - \bar{\alpha}_{t-1} - (\sigma_t^s)^2} \frac{x_t^s - \sqrt{\bar{\alpha}_t}x_t^{s,\text{mix}}}{\sqrt{1 - \bar{\alpha}_t}} + \sigma_t^s z$ 
12:   end for
13:   if  $s < N - 1$  then
14:      $\tilde{x}^{s+1} = \hat{x}_0^s \uparrow^r$ 
15:      $z \sim \mathcal{N}(0, \mathbf{I})$ 
16:      $x_{T[s+1]}^{s+1} = \sqrt{\bar{\alpha}_{T[s+1]}}\tilde{x}^{s+1} + \sqrt{1 - \bar{\alpha}_{T[s+1]}}z$ 
17:   end if
18: end for

```

3.3. Guided generation

To guide the generation by a user-provided loss, we follow the general approach of [3], where the gradient of the loss is added to the predicted clean image in each diffusion step. Here we describe two ways to guide SinDDM generations, one by choosing a region of interest (ROI) in the original image and its desired location in the generated image and one by providing a text prompt.

Generation guided by image ROIs In image-guided generation, the user chooses regions from the training image and marks where they should appear in the generated image. The rest of the image is generated randomly, but coherently with the constrained regions (see Fig. 9). To achieve this effect, we use a simple L^2 loss. Specifically, let x_{target}^s be



Figure 5. **Image generation guided by text.** SinDDM can generate diverse samples guided by a text prompt. The strength of the effect is controlled by the strength parameter η (blue), while the spatial extent of the affected regions is controlled by the fill factor f (orange).

an image containing the desired contents within the target ROIs and let m^s be a binary mask indicating the ROIs, both down-sampled to scale s . Then we define our ROI guidance loss to be $\mathcal{L}_{\text{ROI}} = \|m^s \odot (\hat{x}_0^s - x_{\text{target}}^s)\|^2$. Taking a gradient step on this loss boils down to replacing the current estimate of the clean image, \hat{x}_0^s , by a linear interpolation between \hat{x}_0^s and x_{target}^s . Namely

$$\hat{x}_0^s \leftarrow m^s \odot ((1 - \eta) \hat{x}_0^s + \eta x_{\text{target}}^s) + (1 - m^s) \odot \hat{x}_0^s, \quad (2)$$

where the step size η determines the strength of the effect. We use this guidance in all scales except for the finest one.

Text guided style For text-guidance, we use a pre-trained CLIP model. Specifically, in each diffusion step we use CLIP to measure the discrepancy between our current generated image, \hat{x}_0^s , and the user’s text prompt. We do this by augmenting both the image and the text prompt, as described in [1] (with some additional text augmentations described in App. E.2), and feeding all augmentations into CLIP’s image encoder and text encoder. Our loss, $\mathcal{L}_{\text{CLIP}}$, is the average cosine distance between the augmented text embeddings and the augmented image embeddings. We update \hat{x}_0^s based on the gradient of $\mathcal{L}_{\text{CLIP}}$. At the finest scale $s = N - 1$, we finish the generation process with three diffusion steps without CLIP guidance. Those steps smoothly blend the objects created by CLIP into the generated image. For style guidance, we provide a text prompt of the form “X style” (e.g. “Van Gogh style”) and apply CLIP guidance only at the finest scale. To control the style of random samples, all pyramid levels before that scale generate a random sample as usual and are thus responsible for the global structure of the final sample. To control the style of the training image itself we inject that image directly to the finest scale, so that the modifications imposed by our denoiser and by the CLIP guidance only affect the fine tex-

tures. This leads to a style-transfer effect, but where the style is dictated by a text prompt rather than by an example style image (see Fig. S9 in the appendix).

Text guided contents To control contents using text, we use the same approach as above, but apply the guidance at all scales except $s = 0$. We also constrain the spatial extent of the affected regions by zeroing out all gradients outside a mask m^s . This mask is calculated in the first step CLIP is applied, and is kept fixed for all remaining timesteps and scales (it is upsampled when going up the scales of the pyramid). The mask is taken to be the set of pixels containing the top f -quantile of the values of $\nabla_{\hat{x}_0^s} \mathcal{L}_{\text{CLIP}}$, where $f \in [0, 1]$ is a user-prescribed fill factor. We use an adaptive step size strategy, where we update \hat{x}_0^s as

$$\hat{x}_0^s \leftarrow \eta \delta m^s \odot \nabla \mathcal{L}_{\text{CLIP}} + (1 - m^s) \odot \hat{x}_0^s. \quad (3)$$

Here $\delta = \|\hat{x}_0^s \odot m\| / \|\nabla \mathcal{L}_{\text{CLIP}} \odot m\|$ and $\eta \in [0, 1]$ is a strength parameter that controls the intensity of the CLIP guidance. We also use a momentum on top of this update scheme (see App. E.1). We let the user choose both the fill factor f and the strength η to achieve the desired effect. Their influence is demonstrated in Fig. 5.

4. Experiments

We trained SinDDM on images of different styles, including urban and nature scenery as well as art paintings. We now illustrate its utility in a variety of tasks.

Unconditional image generation As illustrated in Figs. 1, 6, S1 and S15, SinDDM is able to generate diverse, high quality samples of arbitrary dimensions. Close inspection reveals that SinDDM often generalizes beyond the structures appearing in the training image. For example, in Fig. 1, 2nd row, the angles of many of the mountains in the

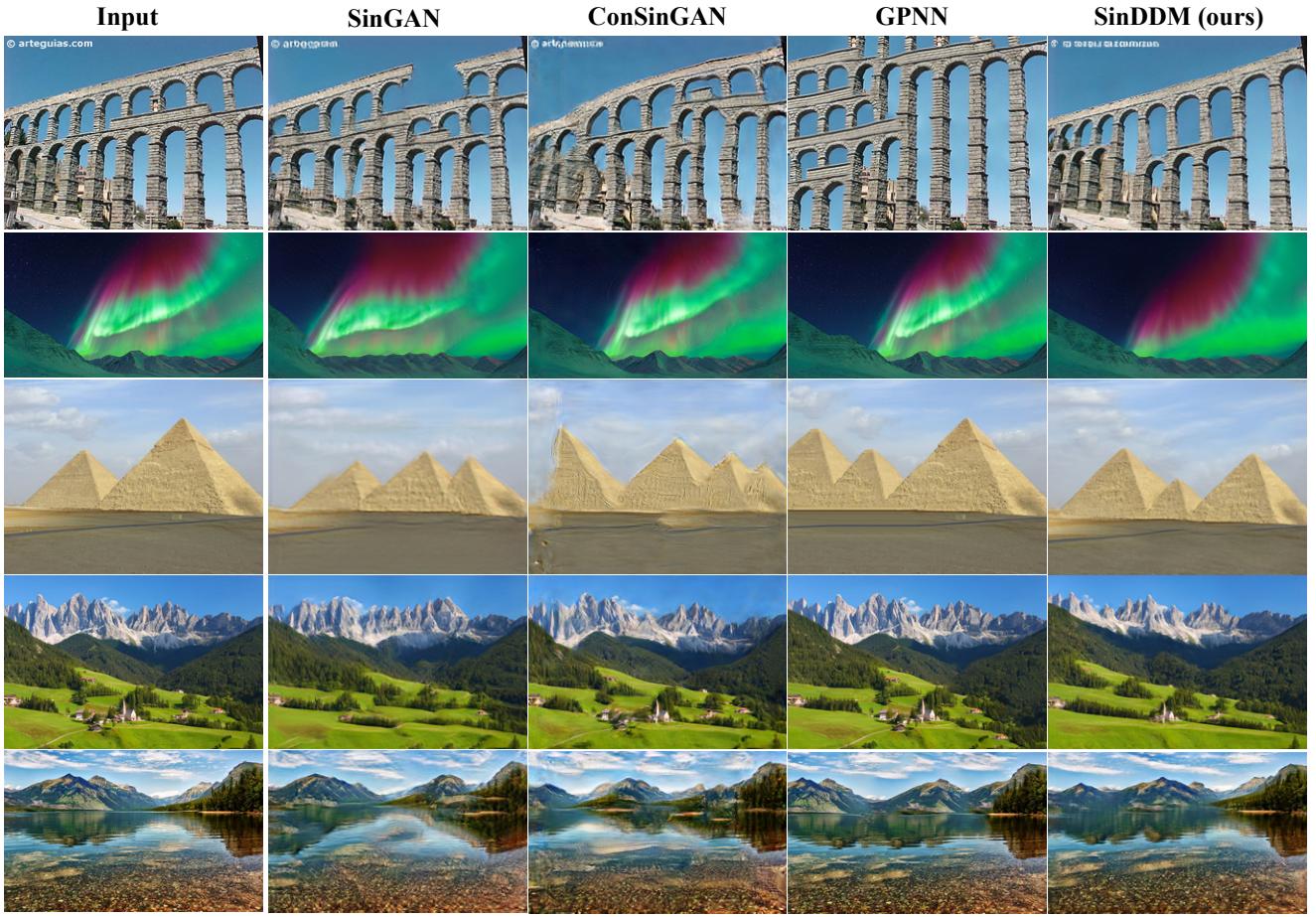


Figure 6. **Unconditional image generation comparisons.** We qualitatively compare our model to other single image generative models on unconditional image generation. As can be seen, our results are at least on par with the other models in terms of quality and generalization.

leftmost sample do not appear in the training image. Table 1 reports a quantitative comparison to other single image generative models on all 12 images appearing in this paper (see App. G.1 for more comparisons). Each measure in the table is computed over 50 samples per training image (we report mean and standard deviation). As can be seen, the diversity of our generated samples (both pixel standard-deviation and average LPIPS distance between pairs of samples) is higher than the competing methods. At the same time, our samples have comparable quality to those of the competing methods, as ranked by the no-reference image quality measures NIQE [15], NIMA [28] and MUSIQ [13]. However, the single-image FID (SIFID) [24] achieved by SinDDM is higher than the competing methods. This is indicative of the fact that SinDDM generalizes beyond the structures in the training image, so that the internal deep-feature distributions are not preserved. Yet, as we show next, this does not prevent from obtaining highly satisfactory results in a wide range of image manipulation tasks.

Metric	SinGAN	ConSinGAN	GPNN	SinDDM
Pixel Div. \uparrow	0.28 ± 0.15	0.25 ± 0.2	0.25 ± 0.2	0.32 ± 0.13
LPIPS Div. \uparrow	0.18 ± 0.07	0.15 ± 0.07	0.1 ± 0.07	0.21 ± 0.08
NIQE \downarrow	7.3 ± 1.5	6.4 ± 0.9	7.7 ± 2.2	7.1 ± 1.9
NIMA \uparrow	5.6 ± 0.5	5.5 ± 0.6	5.6 ± 0.7	5.8 ± 0.6
MUSIQ \uparrow	43 ± 9.1	45.6 ± 9	52.8 ± 10.9	48 ± 9.8
SIFID \downarrow	0.15 ± 0.05	0.09 ± 0.05	0.05 ± 0.04	0.34 ± 0.3

Table 1. Quantitative evaluation for unconditional generation.

Generation with text guided contents Figures 2, 5, S2 present text guided content generation examples. As can be seen, our approach allows obtaining quite significant effects, while also remaining loyal to the internal statistics of the training image. In Figs. 2 and S3 we illustrate editing of local regions via text. In this setting, the user chooses a ROI and a corresponding text prompt. These are used as inputs to CLIP’s image and text encoders, and the gradients of the CLIP loss are used to modify only the ROI. In Figs. 7, and S16-S18 we compare our text-guided content genera-



Figure 7. **Image generation and editing guided by text.** We compare SinDDM to Text2Live and stable diffusion (using the approach of SDEdit). Unlike these methods, SinDDM is not constrained to the aspect ratio or scene arrangement of the training image. We used the text prompts “*stars constellations in the night sky*” and “*volcano eruption*” for the 1st and 2nd rows, respectively. Text2Live requires four different text prompt as inputs. For the pyramid image, we supplied it with the additional texts “*volcano erupt from the pyramids in the desert*” to describe the the full edited image, “*pyramids in the desert*” to describe the input image and “*the pyramids*” to describe the ROI in the input image (see App. G.2 for the text prompts we used for the night sky image). For stable diffusion we tried many strength values and chose the best result (see App. G.2 for other strengths).

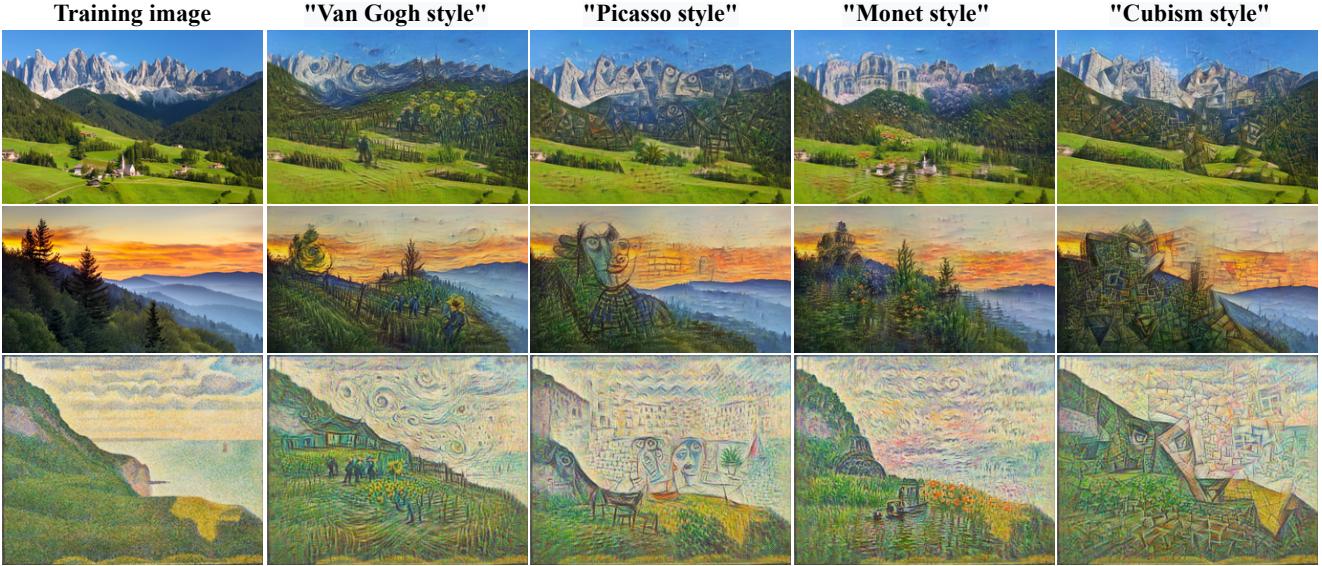


Figure 8. **Generation with text guided style.** SinDDM can generate samples in a prescribed style using CLIP guidance at the finest scale.

tion method to Text2Live [1] and to Stable Diffusion [19]. Text2Live is an image editing method that can operate on any image (or video) using a text prompt. It does so by synthesizing an edit layer on top of the original image. The edit is guided by four different text prompts that describe the input image, the edit layer, the edited image and the ROI. This method cannot move objects, modify scene arrangement, or generate images of different aspect ratios. Our model is guided only by one text prompt that describes the desired result and can generate diverse samples of arbitrary dimensions. As for Stable Diffusion, we use the “image-to-image” option implemented in their source code. In this set-

ting, the image is embedded into a latent space and injected with noise (controlled by a strength parameter). The denoising process is guided by the user’s text prompt, similarly to the framework described in SDEdit [14] (see App. G.2).

Generation with text guided style Figures 2, 8, S4-S8 present examples of image generation with a text-guided style. Here, the guidance generates not only the textures and brush strokes typical of the desired style, it also generates fine semantic details that are commonly seen in paintings of this style (*e.g.* typical scenery, sunflowers in “Van Gogh style”). Figure S9 shows text-guided style transfer.

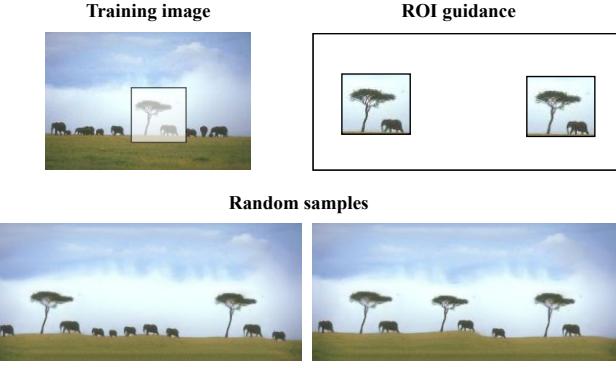


Figure 9. **Image guided generation in ROIs.** Our model is able to generate images with user-prescribed contents within several ROIs. The rest of the image is generated randomly but coherently around those constraints.

Generation guided by image ROIs Figures 9 and S10 show examples for generation guided by image ROIs. Here, the goal is to generate samples while forcing one or more ROIs to contain pre-determined content. SinDDM generates diverse contents outside the constrained ROIs, that coherently stitch with the constrained regions.

Style transfer Similarly to SinGAN, SinDDM can also be used for image manipulation tasks, by relying on the fact that it can only sample images with the internal statistics of the training image. Particularly, to perform style transfer, we train our model on the style image and inject a downsampled version of the content image into some scale $s \leq N - 1$ and timestep $t \leq T$ (by adding noise with the appropriate intensity). We then run the reverse multi-scale diffusion process to obtain a sample. At the injection scale, we use $\gamma_t^s = 0$ for all t and match the histogram of the content image to that of the style image. As can be seen in Figs. 10 and S24, this leads to samples with the global structure of the content image and the textures of the style image. We show a qualitative comparison with SinIR [30], a state-of-the-art internal method for image manipulation.

Harmonization Here, the goal is to realistically blend a pasted object into a background image. To achieve this effect, we train SinDDM on the background image and inject a downsampled version of the naively pasted composite into some scale s and timestep t (with $\gamma_t^s = 0$ at the injection scale). As can be seen in Figs. 11 and S25, SinDDM blends the pasted object into the background, while tailoring its texture to match the background. Here, our result is less blurry than SinIR’s.

5. Conclusion

We presented SinDDM, a new single image generative model that combines the power and flexibility of DDMs

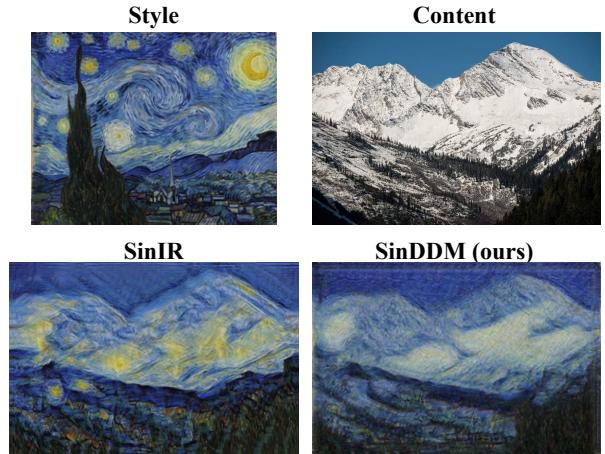


Figure 10. **Style transfer.** SinDDM can transfer the style of the training image to a content image, while preserving the global structure of the content image.

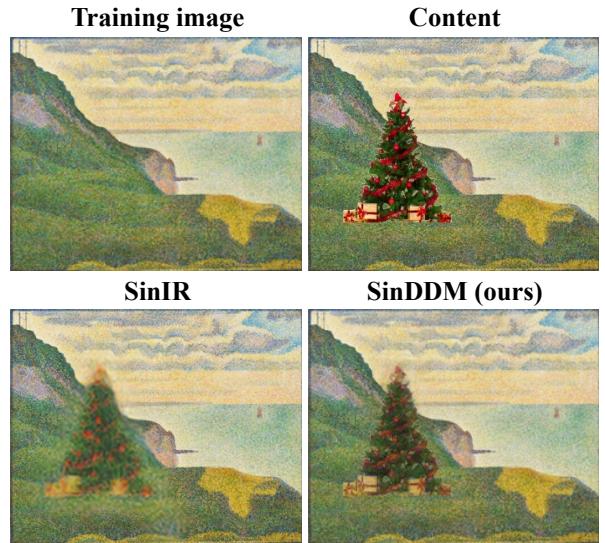


Figure 11. **Harmonization.** Injecting an image with a naively pasted object into an intermediate scale and timestep, matches the object’s appearance to the training image.

with the multi-scale structure of SinGAN. SinDDM can be easily guided by external sources. Particularly, we demonstrated text-guided image generation, where we controlled the contents and style of the samples. A limitation of our method is that it is often less confined to the internal statistics of the training image than other single image generative techniques. While this can be advantageous in tasks like style transfer (see the colors in Fig. 10), in tasks like unconditional image generation, this can lead to over- or under-representation of certain objects in the image (see App. H).

References

- [1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2LIVE: Text-driven layered image and video editing. In *European conference on computer vision*, 2022. [3](#), [5](#), [7](#)
- [2] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European Conference on Computer Vision*, 2022. [1](#)
- [3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. [1](#), [2](#), [3](#), [4](#)
- [4] Ariel Elnekave and Yair Weiss. Generating natural images with direct patch distributions matching. *arXiv preprint arXiv:2203.11862*, 2022. [2](#)
- [5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. [3](#)
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. [1](#)
- [7] Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, and Michal Irani. Drop the GAN: In defense of patches nearest neighbors as single image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13460–13469, 2022. [2](#)
- [8] Gal Greshler, Tamar Shaham, and Tomer Michaeli. Catch-a-waveform: Learning to generate audio from a single short example. *Advances in Neural Information Processing Systems*, 34:20916–20928, 2021. [2](#)
- [9] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch VAE-GAN: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, 33:16761–16772, 2020. [2](#)
- [10] Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. Improved techniques for training single-image GANs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1300–1309, 2021. [2](#)
- [11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. [3](#)
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [1](#), [2](#), [3](#), [4](#)
- [13] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. [6](#)
- [14] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. [1](#), [3](#), [7](#)
- [15] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. [6](#)
- [16] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. [1](#)
- [17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. [2](#), [3](#)
- [18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. [1](#), [3](#)
- [19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. [1](#), [3](#), [7](#)
- [20] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. [1](#), [3](#)
- [21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Ghasemipour, Burcu Ayan, S. Mahdavi, Rapha Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion

- models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 1, 3
- [22] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 3
- [23] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1
- [24] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 2, 6
- [25] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and retargeting the” dna” of a natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4492–4501, 2019. 2
- [26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1, 2
- [27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4
- [28] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 6
- [29] Rundi Wu and Changxi Zheng. Learning to generate 3d shapes from a single example. *ACM Transactions on Graphics (TOG)*, 41(6), 2022. 2
- [30] Jihyeong Yoo and Qifeng Chen. Sinir: Efficient general image manipulation with single image reconstruction. In *International Conference on Machine Learning*, pages 12040–12050. PMLR, 2021. 8
- [31] Zilong Zheng, Jianwen Xie, and Ping Li. Patchwise generative convnet: Training energy-based models from a single natural image for internal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2961–2970, 2021. 2